Goldstein, H., & Leckie, G. (2016). Trends in examination performance and exposure to standardised tests in England and Wales. *British Educational Research Journal*, *42*(3), 367-375. https://doi.org/10.1002/berj.3220

Trends in examination performance and exposure to standardised tests in England and Wales.

Harvey Goldstein and George Leckie
Graduate School of Education
University of Bristol

Abstract

Schools in England and Wales since the late 1980s have been compared in terms of their performances in public examinations and standardised test scores in the form of 'school league tables', with Wales ceasing to produce these after 2001. One of the factors related to performance in examinations is the choice of the examination board, with five main boards currently active in England and Wales. In this paper, we study differences in student uptake and performance among boards and how this has changed over time. By contrasting the experiences of England and Wales the results of our analyses provide a commentary on recent attempts to understand the effects of league tables on school performance.

Keywords

School league tables, examination performance, national comparisons, accountability, value-added

Address for correspondence

Professor Harvey Goldstein
University of Bristol
Graduate School of Education
35 Berkeley Square
Bristol BS8 1JA
h.goldstein@bristol.ac.uk

1. Introduction

There is a continuing debate, in the UK and elsewhere, about whether school league tables based upon standardised test scores and examination results, despite their well-known negative side effects, actually improve the performance of pupils by 'holding schools to account' and by facilitating 'the market in education' via informing choice of schools.

The use of value-added scores, based upon analyses that incorporate measures on individual students, is generally regarded as the soundest basis for appropriate school comparisons since they (attempt to) adjust for selection effects that lead to schools differing in the initial achievements of their students. Nevertheless, Leckie and Goldstein (2009, 2011a, 2011b) argue that the uncertainty surrounding value-added scores for schools is so large that most schools cannot reliably be separated for the purposes of accountability. Furthermore, they conclude that such scores are of even less use for parents choosing schools for their children, where they need to predict school performance many years ahead given past data.

One of the more recent analyses that claims that public league tables do in fact raise pupil performance, by holding them to account, is by Burgess et al. (2013). These authors compare the year 11 (modal age 16 years) General Certificate of Secondary Education (GCSE) public examination scores of students over time in Wales, where league tables were dropped in 2001, and England which continued to publish them. The authors argue that the abolition of league tables in Wales worsened the performance of Welsh students relative to their English peers. They conclude that "If uniform national test results exist, publishing these in a locally comparative format appears to be an extremely cost-effective policy for raising attainment and reducing inequalities in attainment."

In this paper, we first discuss the Burgess et al. findings and examine the assumptions that are made. We then present a new analysis that addresses the same country comparison, using a methodology based upon comparing student performances across different examination boards over time.

2. Trends across time in England and Wales

Essentially, Burgess et al. compare the difference in GCSE results between England and Wales over the period 2002 – 2008 and show that this difference increases steadily over time. This compares with the period before 2002 when there were no differential trends over time. The authors are careful to try to rule out causes other than the abolition of league tables in Wales in 2001 for this trend, testing various assumptions and using a series of carefully

constructed statistical models. Nonetheless, they conclude that it is indeed the abolition of the league tables that has placed Wales at an increasing disadvantage compared to England.

One limitation of these authors' analyses is that they only had data at the school level and any relationships at that level may not reflect those that pertain to relationships of interest which are at the student level. This is often referred to as the 'ecological fallacy' (Robinson, 1950). A second limitation is that they were only able to perform rough approximations to full value-added analyses by using (school average) test score data at key stage 3 (KS3, year 9, modal age 14) when students had already been at secondary school for three years rather than the key stage 2 (KS2, year 6, modal age 11) data based on tests just before starting secondary school. Thus, their analyses only allow for a partial adjustment for secondary school selection mechanisms, restricting their ability to draw causal conclusions. A strength of our analysis below is that we have access to both GCSE results data at the student level and that we can link these data to students' KS2 scores.

A more general concern with studies of this kind, referred to as 'difference-in-differences' studies within economics, is that attempts to infer causation from correlated trends are open to criticism because so many other things also change over the study period and one can never be sure that these alternative explanations have been completely ruled out.

Another issue, and the focus of this paper, concerns the substantial differential use of examination boards across the two countries. In Wales most students sit the Welsh Joint Examination Council (WJEC) exams whereas in England only a minority do. A basic requirement when public examination scores are used for comparisons is that there is an equivalence of marking and grading standards across different boards, or at least in this case that any differences are constant over time. Yet this is a problematic assumption. When the league tables and associated KS2 and KS3 testing were abolished in Wales, there was no longer any satisfactory way that such common tests could be used to establish and monitor exam standards in Wales in relation to England. There is therefore a concern that comparability may have changed over time. For example, the relative Welsh decline in exam results might potentially be explained by a relative hardening of the WJEC examination difficulty relative to that of the English exam boards. Burgess et al. merely state, without providing evidence, that: "The National Qualifications Framework ensured that qualifications attained by pupils across the countries were comparable during this period". One of the ways in which the authors *could* have put this comparability assumption to the test, albeit weakly, while simultaneously going some way to evaluating the robustness of their 'causal' finding, would have been to divide England into its regions and study comparative trends in each of

these, in order to see if in fact Wales really was behaving differently to *all* other regions. If similar variations in trends were to be found among English regions it would then seem less plausible that the Welsh results could be ascribed to changes in either league table or testing policy.

The major omission in the paper, however, is that the authors fail to mention that, by stopping KS2 and KS3 testing at the same time as abolishing league tables, Welsh students became less exposed to high-stakes tests, and were therefore arguably less well-equipped for the GCSE examinations too. This, admittedly, is somewhat speculative, but we do know that the ability to do well on tests is strongly related to the amount of practice that pupils have been given (see for example, Rogers and Yang, 1996), and it would be somewhat surprising if this did not also extend to public examinations. Interestingly, when piloting for the reintroduction of regular testing in Wales took place in 2012, there was evidence (Hodgson, 2013) that the performance of pupils had deteriorated as a result of not being tested intensively during their schooling. So here we have a plausible mechanism that is capable of explaining the relative Welsh decline in exam results. In other words it may have little if anything to do with non-publication of league tables, but simply to the lack of test practice. We elaborate further on this in the discussion.

We cannot, with the data available, hope to resolve the 'causality' issue, but we can explore further the problem of comparability between examination boards and the remainder of this paper sets out to do that. In addition the paper explores the role of school's exam board entry policies in terms of students' KS2 scores. In the next section we study overall trends in uptake and performance in each country and by exam board. We then look at trends in value-added exam performance, where we fit multilevel statistical models to adjust students' GCSE scores for their KS2 performances, to facilitate fairer and more meaningful comparisons between exam boards. This of course necessitates restricting our analysis to English schools due to the stopping of KS2 testing in Wales. We are therefore forced to contrast the value-added exam performance of English students who take the WJEC exams to English students who do not.

3. Exam board differences in GCSE English in England, 2007 – 2011

Burgess et al. analyse the GCSE exam results for the period 2004-2008. In our analyses we have student-level data, including students' KS2 results, for the period 2007-2011 and so our focus is beyond the period studied by those authors.

Our data are drawn from the National Pupil Database (NPD) maintained by the UK Department for Education (https://www.gov.uk/government/collections/national-pupil-database). The data contain GCSE examination grades and boards in all subjects taken by all pupils in England together with their average test scores (across English and mathematics) at KS2 when they are in their last year of primary schooling. The data are restricted to pupils in maintained schools at both occasions. Over this period approximately six percent of schools changed their formal status to academies or free schools, as a result of Government policies designed to increase school autonomy. We have considered such schools as being the same school over the period 2007-2011. The two most important subjects for school league tables are English language and mathematics. In England over this period, only 0.25% of those taking mathematics exams sat those set by the WJEC, in contrast to 17% for English language. We therefore only study results for the latter. The three other exam boards used by schools for this subject are: Assessment and Qualifications Alliance (AQA); Oxford, Cambridge and Royal society of Arts (OCR); and Pearson Education Limited (Pearson). We analyse examination grade scores computed using the standard point scoring system as follows: A* = 58, A = 52, B = 46, C = 40, D = 34, E = 28, F = 22, G = 16, U = 0. Over the five year period the total sample for analysis is 2,740,182 students.

Table 1 and Figure 1 show the numbers of pupils involved by examination board and year together with mean grade scores. Apart from the overall increase in mean scores, one trend that stands out from this table is the rapid increase in the numbers of students in England taking the WJEC examination, whereas the three other exam boards lose numbers We have no direct evidence about why this increase occurred, but possibly it was partly because in schools that had not entered students for WJEC examinations previously, WJEC was increasingly perceived to be easier over our period, and that this also encouraged schools to enter lower achieving students for WJEC examinations. It could also have occurred partly as a result of marketing activity by the WJEC board. By contrast, for 'WJEC schools' – the subset of schools who have at least one WJEC candidate in every year they appear in the data (some schools opened and closed during the period) – the overall numbers remain stable and their overall mean scores are similar to the remainder. We shall look at these schools below. The other trend is the fall in performance of the WJEC students in Wales compared to those in England (including WJEC students in England). This is consistent with the findings of Burgess et al. for 2004-2008, although they looked at mean GCSE score over all subjects, not just English. We explore this issue below in our more detailed multilevel analysis.

Table 2 shows a series of three-level random-intercept models (Leckie, 2013; Goldstein, 2011) where the response is student GCSE grade score and predictors are year (centered on 2007 and treated as a continuous variable), whether or not the exam board was WJEC, and the student's average KS2 test score (across English and mathematics), plus selected interaction terms. These models take account of the fact that both schools and local authorities differ in terms of GCSE achievement. Each school and each authority contributes an additional 'effect', positive or negative, to GCSE performance and the multilevel model incorporates these and provides an estimate of the between-school and between-authority variance of the effects. It is important to account for such 'clustering' in statistical models as failure to do so will typically lead to spuriously precise regression coefficients and thus incorrect inferences. The three-level structure of the models therefore take into account the nesting of students (level-1) within schools (level-2) within local authorities (level-3).

There are differences in GCSE grade scores among the three English exam boards, but for simplicity we group these in each model in order to focus on the overall differences between these and the WJEC board.

Model A shows an average lower grade score in 2007 of 0.716 for those taking WJEC with a greater rate of increase of 0.048 over time for this group. Thus, students in English schools sitting the WJEC exams show the opposite pattern (i.e., convergence over time) from those in Wales (i.e., divergence over time) as given in Table 1. The estimated variance components confirm that we need the multilevel approach: there is modest residual clustering at the LA-level (2% of the variation lies between LAs; 0.020 = 1.886/(1.886 + 16.953 + 77.216)) and substantial residual clustering of 18% at the school-level.

Model B extends this analysis by adjusting for student KS2 score so that the remaining coefficients are to be interpreted in terms of student progress during secondary school – a value-added analysis. The gap in 2007 between WJEC and other students is now smaller reflecting the lower KS2 achievement of the WJEC students. The exam board attainment gap, however, shows a similar rate of narrowing over time as before (0.049): over the four year period, the difference between WJEC and remaining students decreases somewhat by 0.196 (= 4*0.049) grade score points. We see, in the interaction term with KS2, a greater coefficient for KS2 among the WJEC so that the GCSE attainment gap between WJEC and non-WJEC students narrows with higher performance on KS2. Thus, the WJEC student mean performance for students at the lower 10th percentile on KS2 score (a KS2 score of -5.920) is -0.720 (= -0.347 + 0.063*-5.920) that is, over two thirds of a grade score behind whilst for those at the 90th

percentile on KS2 score (a KS2 score of 5.620) it is -0.007 (= -0.347 + 0.063*5.620) that is effectively the same as the non-WJEC student mean performance.

In Models C and D we have included terms that distinguish 'WJEC schools', which entered at least one WJEC student in every year that they appear in the data, from 'non-WJEC schools' which did not. In Model C, we see that there is a non-significant effect for being in a WJEC school in 2007 but a significant decreasing score over time so that by 2011 students in these schools are 0.427 (= -0.039 – 0.097*4) grade scores apart. From Model D, we see that the predicted value in 2007 between WJEC students in schools with some WJEC candidates in every year compared to non-WJEC students in the remaining schools is -0.565 (= -0.945 - 0.591 + 0.971) and this is relatively unchanged by 2011 becoming -0.489 (= -0.945 + 0.248*4 -0.591 + 0.214*4 + 0.971 - 0.443*4). When compared with the marked decrease of 2.6 (= 35.4 - 38.0) grade scores in Welsh WJEC students' grades over the period (Table 1), this reinforces the inference of a real decline in scores for Welsh students.

It is of interest to look at the exam board entry policies of schools, especially for whether there are selection effects with respect to pupil characteristics. Table 3 presents the results of Models E and F which looks at the relationship between KS2 score and school exam board entry policy. In Model E, we see that overall there is a lower average KS2 in the WJEC schools (-0.300). When, however, in Model F we actually look at those students who take the WJEC exams, we see only a small and non-significant interaction between taking WJEC and being in a WJEC school (-0.088), and that it is the students who take the WJEC board examination who really have the lower KS2 scores. The non-WJEC students in WJEC schools actually scored 0.301 points higher at KS2 than the non WJEC students in non-WJEC schools. Thus, this suggests that the WJEC schools have a lower achieving intake and are entering 'lower achieving' students for WJEC than for other exam boards at GCSE.

In summary, therefore, it would seem that any change in comparability of grading standards between WJEC and other boards is not a major explanation for the finding of decreasing performance of pupils in Wales.

4. Conclusions

Our conclusion from these analyses is that there does appear to be a real decline in Welsh pupil test scores over 2007-2011 following both the abolition of regular testing in schools and abolition of school league tables. Since our analyses are able to adjust for prior achievement in the form of KS2 tests our conclusions reflect the temporal changes in progress that students make during secondary schooling. Our analysis reveals some other interesting

features. Thus, the schools that have WJEC students throughout the period tend to have a lower achieving intake and are entering 'lower achieving' students for WJEC than for other exam boards at GCSE. This is consistent with the view that such students are perceived to perform better if entered for WJEC exams.

Attribution of any causal effects is not possible given the data available and, in particular, the inference that the abolition of league tables per se is responsible is not supported by evidence. In our view it seems more reasonable to suppose that a lack of test practice ('test-wiseness', Rogers and Yang, 1996) is a major contributing factor. This is reinforced by a study carried out in 2012 which also showed that Welsh pupils who had not been exposed to regular standardised tests tended, for example, to lack even the ability to navigate successfully through test booklets (Hodgson, 2013).

If this inference is accepted it suggests also that a lack of test practice on key stage tests either has a direct effect on GCSE performance or an indirect effect, for example through less attention being given to preparation for GCSE. Possibly both of these pathways might operate. Our data allow no further insight into such possibilities. The presence of league tables may provide an incentive for large amounts of test practice, but it does not follow that league tables are necessary to ensure adequate preparation for GCSE, especially in the light of their now well-understood negative side effects and perverse incentives (Foley and Goldstein, 2012).

The principal purpose of the present paper is to investigate whether the issue of exam board comparability could be a major explanation for the observed effect. Our analyses suggest that this is not so.

Given the retrospective nature of the data available for research we do not envisage that it can be made to yield more causally robust inferences. With hindsight, it would have been useful when the Welsh test reduction policy was introduced, to have built in an evaluation of its effects, both in terms of examination results and performance on international tests such as those conducted by the Programme for International Student Assessment (PISA). A randomised controlled trial would have been difficult to implement, in particular establishing a valid control group. Nevertheless, a careful monitoring of school and teacher testing behaviour could have been implemented and this would have yielded useful information and even alerted teachers to the possibility that GCSE preparation may have been affected. At the same time the effects of abolishing Key Stage tests on the curriculum could have been studied as well as the effects on student and teacher motivations. We suggest that this

discussion of Welsh student performance serves to emphasise the general importance of independent monitoring of major educational innovations from the outset.

Acknowledgements

We are grateful to the Department for Education (UK) for access to the National Pupil Database. Leckie was funded by UK Economic and Social Research Council grant ES/K000950/1.

References

Burgess, S., Wilson, D., and Worth, J. (2013). A natural experiment in school accountability: The impact of school performance information on pupil progress. *Journal of Public Economics*, 106, 57–67.

Foley, B., and Goldstein, H. (2012). *Measuring Success: league tables in the public sector*. London, British Academy.

Goldstein, H. (2011). *Multilevel Statistical Models*, 4th edition. Chichester, UK: Wiley.

Hodgson, C. (2013). Eales – gone full circle? A brief history of assessment in Wales and what is happening now. Paper presented at 14th AEA-Europe Annual Conference, Paris, 7-9 November 2013.

Leckie, G. (2013). Three-Level Multilevel Models - Concepts. LEMMA VLE Module 11, 1-47.

Leckie, G. and Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 835-851.

Leckie, G. and Goldstein, H. (2011a). A note on "The limitations of using school league tables to inform school choice". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 833-836.

Leckie, G. and Goldstein, H. (2011b). Understanding uncertainty in school league tables. *Fiscal Studies*, 32, 207-224.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.

Rogers, W.T. and Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, 12, 247-259.

Table 1. Mean GCSE English exam grade scores by examination board and year. England only. Numbers of pupils in brackets.

	2007	2008	2009	2010	2011
AQA	38.9 (426,183)	39.1 (421,545)	39.5 (406,811)	40.3 (401,986)	40.9 (390,355)
OCR	38.4 (39,967)	38.5 (32,539)	39.2 (28,105)	40.2 (23,634)	41.0 (21,718)
Pearson	38.9 (24,469)	40.3 (21,008)	40.7 (18,169)	41.3 (17,837)	42.2 (17,127)
WJEC	38.1 (70,768)	38.8 (82,567)	39.1 (89,005)	39.4 (100,234)	39.8 (106,155)
Total	38.8 (561,387)	39.1 (557,659)	39.5 (542,090)	40.2 (543,691)	40.7 (535,355)
WJEC school *	38.1 (75,570)	38.8 (74,142)	39.1 (72,717)	39.6 (73,814)	40.2 (73,602)
WJEC Wales **	38.0 (45,908)	38.1 (45,464)	37.1 (45,559)	36.9 (49,204)	35.4 (37,971)

Note: Examination board abbreviations: AQA = Assessment and Qualifications Alliance; OCR = Oxford, Cambridge and Royal society of Arts; Pearson = Pearson Education Limited; WJEC = Welsh Joint Examination Council. * These are schools who have at least one WJEC English candidate in every year they appear in the data. ** These are estimates obtained from tables supplied by the WJEC for all candidates, with students in England removed, using the data in the WJEC row in Table 1. See: http://www.wjec.co.uk/students/results-and-research/results-statistics.html

Table 2. GCSE English exam grade score related to year, WJEC exam board, and KS2 scores. Standard errors in brackets.

	Model A	Model B	Model C	Model D
Intercept	38.467 (0.140)	38.772 (0.132)	38.805 (0.134)	38.812 (0.134)
WJEC	-0.716 (0.048)	-0.347 (0.044)	-0.553 (0.064)	-0.945 (0.083)
Year	0.442 (0.004)	0.392 (0.004)	0.393 (0.004)	0.391 (0.004)
KS2 score		0.550 (0.001)	0.550 (0.001)	0.550 (0.001)
Year × WJEC	0.048 (0.011)	0.049 (0.010)	0.130 (0.020)	0.248 (0.025)
KS2 score × WJEC		0.063 (0.002)	0.063 (0.002)	0.063 (0.002)
WJEC school			-0.039 (0.205)	-0.591 (0.217)
WJEC school × Year			-0.097 (0.021)	0.214 (0.049)
WJEC school × WJEC				0.971 (0.131)
WJEC school × WJEC × Year				-0.443 (0.056)
LA variance	1.886 (0.334)	1.895 (0.300)	1.877 (0.298)	1.877 (0.298)
School variance	16.953 (0.439)	11.286 (0.293)	11.288 (0.293)	11.285 (0.293)
Student variance	77.216 (0.066)	64.400 (0.055)	64.399 (0.055)	64.398 (0.055)
-2 × log likelihood	19703248	19205724	19205250	19205176

Note: WJEC is 1 if WJEC, else 0. Year is centered on 2007. KS2 is average (across English and mathematics) KS2 point score centered on mean of 26.5. WJEC school is 1 if the school has at least one WJEC English entry in every year it appears in the period 2007-2011, else 0.

Table 3. KS2 score (centered) related to year, exam board and school entry policy. Standard errors in brackets.

	Model E	Model F
Intercept	-0.595 (0.093)	-0.592 (0.093)
WJEC		-0.602 (0.031)
Year	0.078 (0.003)	0.089 (0.003)
WJEC school	-0.300 (0.118)	0.301 (0.134)
WJEC × WJEC school		-0.088 (0.074)
LA variance	0.985 (0.144)	1.000 (0.145)
School variance	4.046 (0.106)	4.036 (0.106)
Student variance	41.023 (0.035)	41.016 (0.035)
-2 × log likelihood	17967756	17967282

Note: WJEC is 1 if WJEC, else 0. Year is centered on 2007. KS2 is average (across English and mathematics) KS2 point score centered on mean of 26.5. WJEC school is 1 if the school has at least one WJEC English entry in every year it appears in the period 2007-2011, else 0.

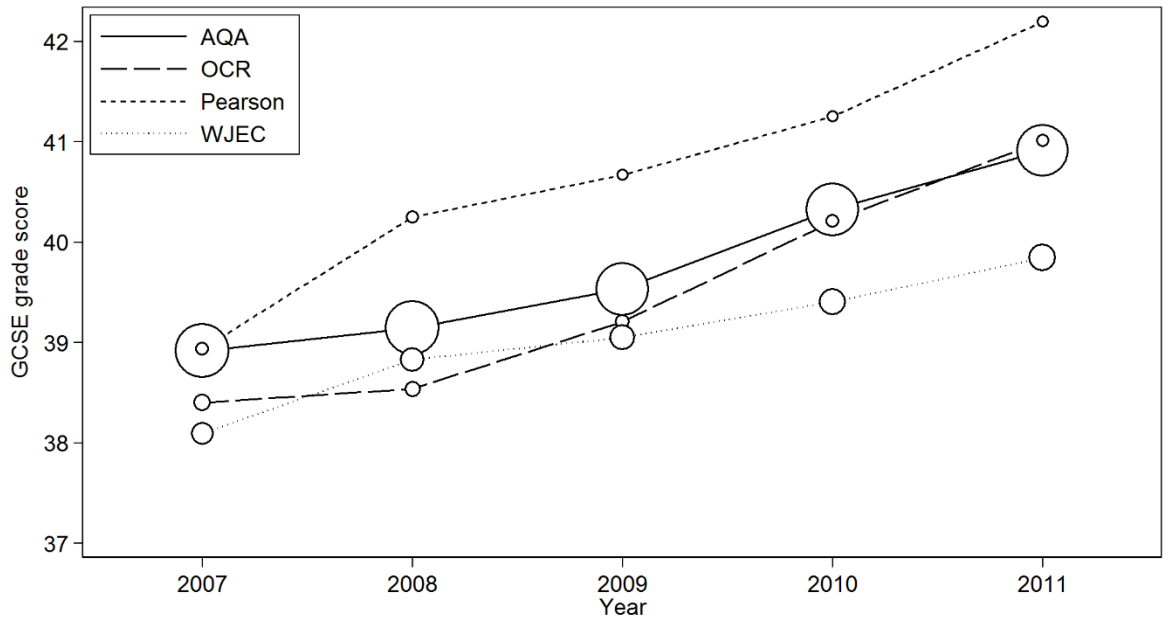

Figure 1. Mean English GCSE grade scores over time plotted separately by exam board. The size of the scatter points are proportional to the number of students sitting the associated exam board in that year.